Assignment-Based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

There are a couple of categorical variables namely season,mnth,yr,weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable ‘cnt’. The below fig shows the correlation among the same These variables are visualized using bar plot and Box plot both.

2. Why is it important to use drop_first=True during dummy variable creation?

The intention behind the dummy variable is that for a categorical variable with ‘n’ levels, you create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables. Eg: If there are 3 levels, the drop_first will drop the first column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The ‘temp’ and ‘atemp’ variables have the highest correlation when compared to the rest with the target variable as ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Linear Regression models are validated based on Linearity,No auto-correlation,Normality of error,Homoscedasticity, Multicollinearity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature,year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail ?

Linear regression is a machine learning algorithm used to model the relationship between a dependent variable (the thing you want to predict) and one or more independent variables (the things you use to make the prediction).
The goal of linear regression is to find the "best-fit" straight line that represents this relationship.

This line is described by an equation in the form:

y = a + bx

Where:
y is the dependent variable you want to predict
x is the independent variable(s) you are using
a is the y-intercept (the value of y when x is 0)
b is the slope of the line (the change in y per unit change in x)

The algorithm works by adjusting the values of a and b to minimize the overall difference between the predicted values (from the line) and the actual observed values in the data.
Once the line is fitted, you can use it to make predictions - if you know the value of the independent variable(s) x, you can plug it into the equation to estimate the corresponding value of the dependent variable y.

The key advantages of linear regression are its simplicity, interpretability, and ability to quantify the relationship between variables. However, it assumes a linear relationship and can be sensitive to outliers or violations of its underlying assumptions.

In summary, linear regression is a fundamental machine learning technique that models the linear relationship between variables to enable prediction and analysis. The algorithm finds the best-fit straight line to represent this relationship.

2. Explain the Anscombe’s quartet in detail?

Anscombe's Quartet is a set of four datasets created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. Despite having nearly identical simple descriptive statistics, the datasets exhibit distinct patterns when graphed, emphasizing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

Characteristics of Anscombe's Quartet:

Dataset Composition: Each dataset in Anscombe's Quartet consists of 11 pairs of x-y values, totaling 44 data points across the four datasets.
Statistical Properties: The datasets have the same mean, variance, correlation coefficient, and linear regression line, showcasing identical summary statistics.
Graphical Representation: When plotted, the datasets reveal unique relationships between x and y, with varying patterns such as linear, non-linear, outliers, and high-leverage points.

Significance of Anscombe's Quartet:

Importance of Data Visualization: Anscombe's Quartet highlights the necessity of visualizing data to uncover nuances, outliers, and diverse relationships that may not be apparent from summary statistics alone.
Challenges Assumptions: By demonstrating how different datasets can lead to misleading conclusions when only numerical summaries are considered, the quartet encourages critical thinking and questioning of assumptions.
Enhances Statistical Literacy: It serves as a valuable tool for teaching statistical concepts, engaging learners in hands-on activities to explore mean, variance, correlation, and regression analysis.
Validates Data and Promotes Discussions: Practitioners can use Anscombe's Quartet to assess the accuracy of datasets, foster open discussions within communities, and drive informed decision-making based on visual and statistical analysis.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables X and Y. It is a number between -1 and 1 that indicates the strength and direction of the relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a data preprocessing technique used to standardize the range of independent variables or features of data. It is an important step in many machine learning algorithms, as it can significantly impact the performance and convergence of the model

The main reasons for scaling data are:

- To avoid attributes in greater numeric ranges dominating those in smaller numeric ranges.
- To ensure faster convergence of gradient-based optimization algorithms used in training machine learning models.
- To prevent numerical difficulties during the calculation.

The key differences between normalization and standardization are:

- Normalization scales the data to a common range (typically 0 to 1), while standardization does not have a predefined range.
- Normalization is more affected by outliers, while standardization is less sensitive to outliers.
- Normalization is useful when the data follows a uniform distribution, while standardization is more appropriate when the data follows a Gaussian distribution.
- Normalization transforms the data into a unit hypercube, while standardization translates the data to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The value of VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity between the independent variables in a regression model. This means that one independent variable can be perfectly predicted by a linear combination of the other independent variables, causing the VIF to become infinite.Perfect multicollinearity occurs when one independent variable can be perfectly predicted by a linear combination of the other independent variables. In other words, there is an exact linear relationship between two or more predictors.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q (Quantile-Quantile) plot is a graphical technique used to determine if two data sets come from populations with a common distribution. It is a scatter plot that compares the quantiles of the empirical distribution of a variable to the quantiles of a theoretical distribution, such as the normal distribution

The key uses and importance of a Q-Q plot in linear regression are:

- Assessing Normality of Residuals: In linear regression, the Q-Q plot is commonly used to check if the residuals (the differences between the observed and predicted values) follow a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot should fall approximately along a straight line.
- Identifying Non-Normality: If the points in the Q-Q plot deviate systematically from the straight line, it suggests that the residuals do not follow a normal distribution. This could indicate issues such as skewness, heavy tails, or the presence of outliers in the data.
- Validating Model Assumptions: The normality of residuals is a key assumption for many statistical inference procedures in linear regression, such as hypothesis testing and confidence interval estimation. The Q-Q plot helps validate this assumption.
- Guiding Data Transformations: If the Q-Q plot reveals non-normality, it can guide the choice of appropriate data transformations (e.g., log transformation, Box-Cox transformation) to improve the normality of the residuals.